Topological Exploration of Embedded GitHub Repository Data Using Mapper

Patrik Zajec
patrik.zajec@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Ivo Hrib
ivo.hrib@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

We present a preliminary topological exploration of a dataset of 10,000 GitHub repositories embedded in a 768-dimensional space. Using the Mapper algorithm from topological data analysis (TDA), we visualize multiple Mapper graphs to reveal branching structures and detect potential cycles. One graph exhibits a cycle corroborated by the persistent homology of the dataset. We also include a 2D UMAP projection for comparison. While formal significance testing is deferred to future work, these visualizations provide insight into latent structural patterns and guide further topological analysis.

Keywords

topological data analysis, Mapper, GitHub, embeddings, visualization, high-dimensional data, persistent homology

1 Introduction

High-dimensional embeddings of software repositories provide rich semantic, structural, and metadata information but remain challenging to interpret. Topological data analysis (TDA) offers tools to summarize the "shape" of data, capturing clusters, branches, and cycles beyond conventional statistics.

The Mapper algorithm is a popular TDA method that constructs a graph based on filter functions, clustering, and overlapping covers. Mapper nodes represent clusters of points, and edges connect overlapping clusters. This approach allows intuitive visualization of high-dimensional data.

In this study, we explore 10,000 GitHub repository embeddings in 768 dimensions. We generate five Mapper graphs to examine branching and potential cycles, compare with a 2D UMAP projection, and analyze persistent homology to corroborate observed cycles. Significance testing of these structures is planned for future work.

2 Related Work

The primary relevant work for this study is:

- **Škraba et al** [1]: Introduces weak universality for significance testing in topological features. While computationally intensive for large datasets, it provides a framework for assessing whether Mapper branches and cycles are meaningful.
- **Patrik Zajec** [2]: Demonstrates Mapper optimization and significance-based pruning of edges. This work serves as the basis for potential future analyses of our observed structures.

Information Society 2025, Ljubljana, Slovenia

3 Dataset and Preprocessing

We use 10,000 GitHub repositories represented as 768-dimensional embeddings from a pretrained transformer. No dimensionality reduction is applied, preserving all original features. Preprocessing involves standard normalization to zero mean and unit variance.

For illustration, a 2D UMAP projection is computed. Although UMAP simplifies the data, it provides intuitive visualization of clustering and branching, complementing Mapper graphs.

4 Methodology

4.1 Mapper Graph Construction

We construct five Mapper graphs using overlapping intervals over the identity filter on a chosen embedding dimension. Within each interval, HDBSCAN clusters the points. Nodes correspond to clusters, and edges represent overlaps. Node attributes such as size and mean filter value are recorded for visualization.

4.2 Persistent Homology

To confirm observed cycles, we compute 1-dimensional persistent homology of the full dataset. Cycles in the Mapper graph that correspond to long-lived features in the persistence diagram indicate potential nontrivial topological structure.

4.3 UMAP Visualization

A 2D UMAP projection of the embeddings is generated for comparison. While the UMAP projection is primarily illustrative and not topologically rigorous, it provides intuition about clusters and branching patterns observed in Mapper graphs. However in our case we were able to gather more data using mapper graphs than our preliminary umap projections. These remain to be verified for validity as mapper graphs can be prone to noise. For this purpose we aim to use the significance testing architecture recently developed, as per [2] based upon [1] in the future.

5 Observations

5.1 Branching Patterns

All five Mapper graphs consistently exhibit branching structure. The branches likely correspond to subsets of repositories with similar characteristics (e.g., programming language or project type). The repetition across multiple graphs suggests robustness to different cover and clustering parameters.

5.2 Cycle Detection

One Mapper graph shows a clear cycle connecting several nodes. The corresponding 1-dimensional persistent homology of the dataset confirms this feature, indicating the presence of a nontrivial loop in high-dimensional embedding space.

5.3 UMAP Patterns

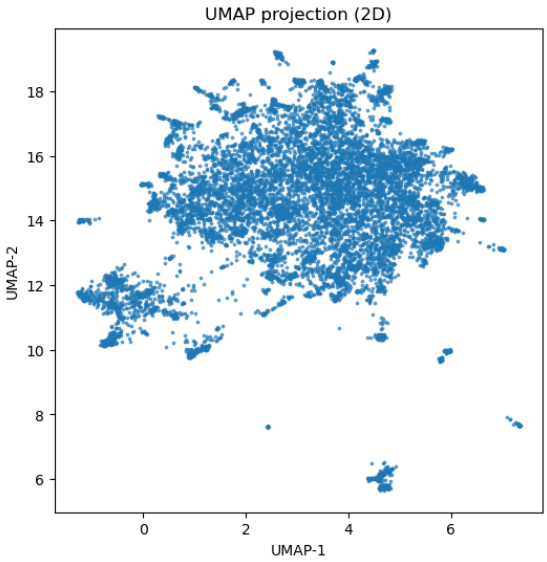

Figure 1: 2D UMAP projection of GitHub embeddings.

The 2D UMAP projection shows no branching patterns that correspond to those seen in Mapper graphs. This projection serves as an intuitive visual check, although it does not provide one in this case, as we are able to discern more about the data's structure topologically

6 Figures and Visuals

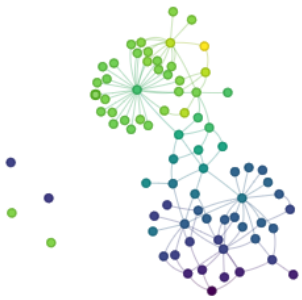

Figure 2: Mapper graph 1 showing some branching structure.Resolution=20 , Overlap=0.2

Figure 3: Mapper graph 2 showing consistent branching. Resolution=20 , Overlap=0.15

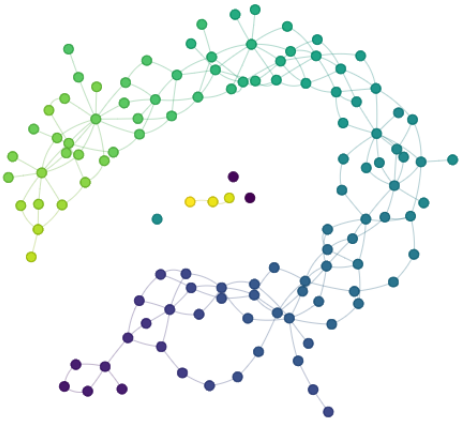

Figure 4: Mapper graph 3 showing similar branching patterns, as well as a more accentuated disconnected component. Resolution=35 , Overlap=0.1

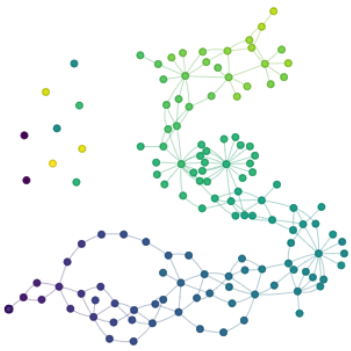

Figure 5: Mapper graph 4 illustrating initial structure of cycle. Resolution=25 , Overlap=0.1

Figure 6: Mapper graph 5 showing branching and a visible cycle. Resolution=40 , Overlap=0.15

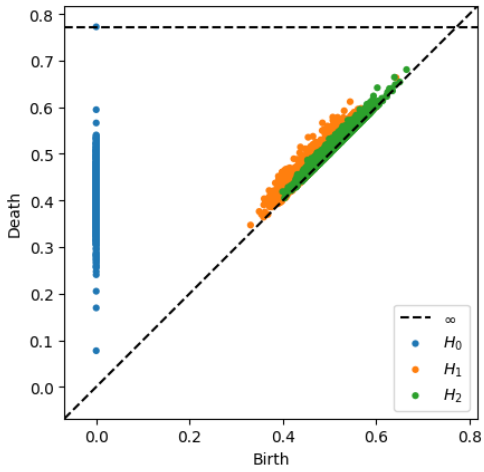

Figure 7: Persistence diagram showing a 1D cycle

7 Discussion and Future Work

Preliminary visual inspection reveals robust branching in all Mapper graphs and a single cycle consistent with persistent homology. While formal significance testing is deferred, these observations suggest latent structure in the high-dimensional embeddings. Future work will include:

- Applying weak universality-based significance testing to Mapper nodes and edges.
- Exploring multiple filter functions and their effect on branching and cycles.
- Correlating topological features with repository metadata (e.g., language, project type).
- Scaling to larger datasets with GPU-accelerated persistent homology computations.

8 Conclusion

We presented a topological exploration of 10,000 GitHub repository embeddings using Mapper and persistent homology. Five Mapper graphs reveal consistent branching, and one graph shows a cycle supported by persistent homology. These observations motivate further investigation into the significance of the identified structures and their semantic interpretation in software repositories. Although preliminary, we aim to expand this approach using weak universality to further interpret and explore relevancy of the structures.

References

[1] Omer Bobrowski and Primož Skraba. [n. d.] A universal null-distribution for topological data analysis. (). https://www.nature.com/articles/s41598-023-37842-2.

[2] Patrik Zajec. 2023. Towards testing the significance of branching points and cycles in mapper graphs. (2023).